International Journal of Distributed Systems and Technologies
Volume 14 • Issue 2

Kinect Body Sensor Technology-Based Quantitative Assessment Method for Basketball Teaching

Youyang Wang, Jilin Institute of Physical Education, China*

ABSTRACT

Emphasizing the process and neglecting the end is the core idea of the research and implementation of college physical education learning and assessment, while the performance is the main form of evaluation results. This paper takes the quantitative assessment of basketball teaching as an example and proposes a new Kinect body sensor technology-based quantitative assessment method for basketball teaching. Specifically, for basketball technology recognition and assessment tasks, the Kinect body sensor is first used to collect volunteer's 3D skeleton motion data, then feeding the collected skeleton sequence to the vision transformer network to model the long-distance dependency. And based on this, the skeleton motion recognition network and skeleton motion assessment network are developed. The experimental results show that the proposed networks can well recognize and quantitatively assess the standard and non-standard basketball skill motions.

KEYWORDS

Artificial Intelligence, Basketball Motion Assessment, Basketball Motion Recognition, Kinect, Skeleton Data, Vision Transformer

1. INTRODUCTION

Emphasizing the process and neglecting the end is the core idea of the research and implementation of college physical education learning and assessment, while the performance is the main form of assessment results. At present, basketball performance assessment usually consists of two parts, i.e., result assessment and skill assessment. The result assessment mainly includes the number of shots in one minute, half court sports shots, etc. And its assessment process is objective and fair, and the evaluators do not need much professional knowledge. The skill assessment is mainly conducted by professionals according to the skill realization process of testers, which including whether the actions are standard and correct. Compared with the result assessment, the technical assessment is relatively independent and flexible, but it is also very dependent on the professional quality of the evaluators, and is also vulnerable to subjective factors. In addition, the continuation of the COVID-19 epidemic

DOI: 10.4018/IJDST.317935 *Corresponding Author

may require Internet teaching and assessment, while online assessment further increases the difficulty of skill assessment. Therefore, it is more and more urgent to research artificial intelligence algorithms that can replace teachers in skill assessment, including feature extraction & feature selection of data analysis (Zheng et al. 2018; Zheng et al. 2022; Zhu et al. 2022), classification of machine learning (Gao et al. 2022; Zhu et al. 2021) etc.

At present, there are many methods for automatic motion recognition, which can be divided into two types: contact type and non-contact type. One branch of the contact type is based on wearable sensors for motion recognition and analysis. However, users using wearable sensors will have a sense of bondage and its cost is very high (Wu et al. 2018). At the same time, this recognition system is vulnerable to external factors to reduce the recognition accuracy. The non-contact motion recognition type mainly uses visual sensors such as cameras to collect data, and uses machine vision and image processing methods to achieve motion recognition and assessment. At present, vision based human motion/posture recognition has become an important topic, and has been widely used in human-computer interaction, virtual reality, intelligent video surveillance and other fields. However, there are still many problems in vision based methods that have not been well solved, which will affect the computer's understanding of human behavior. To be specific, ordinary cameras can only obtain two-dimensional images, but the reconstruction of two-dimensional information to three-dimensional information will lose a lot of important data, affecting the accuracy of motion recognition. Although researchers have designed a variety of image reconstruction algorithms, they are still unable to avoid the effects of lighting, texture occlusion, etc. Kinect sensor uses a new way to obtain images. It captures depth images with spatial distance through a pair of infrared cameras, and extracts skeleton data stream containing three-dimensional coordinate information on the basis of depth images. Therefore, the Kinect body sensor based motion recognition method has gradually attracted extensive attention of researchers.

Kinect is a new generation of somatosensory interactive device that integrates many advanced visual technologies, and it has been highly concerned by many players and researchers in the game and academic circles. Kinect is roughly composed of three parts, RGB camera, depth/infrared camera and infrared sensor, where RGB camera can capture standard two-dimensional color image data stream, and its maximum resolution can reach 1920×1080, the speed can reach 30 fps. The depth camera with infrared ray can capture the depth data stream, with a measurement range of 0.5~4.5 m and a resolution of 512×424, up to 30 fps. Besides, Kinect can obtain three types of raw data: color image data stream, depth data stream and audio data stream, which correspond to three processing processes: identity recognition, skeleton tracking and speech recognition. However, Kinect does not provide advanced functions for motion/pose recognition, because human movements are so changeable that it is difficult to build a universal model for recognition. In addition, although Kinect sensor can perform basic target object motion capture and speech recognition functions, these functions are low-level and the technology is not perfect. Therefore, it is difficult to directly use Kinect's built-in functions to realize basketball motion recognition and assessment tasks.

To solve the above problems, this paper proposes a novel Kinect body sensor technology based quantitative assessment method for basketball teaching. More specifically, we first built a basketball skeleton motion dataset based on Kinect body sensor, and then introduced the vision Transformer (ViT) (Vaswani et al. 2017) network into this paper to develop the ViT based skeleton motion recognition network and ViT based skeleton motion assessment network, respectively, for achieving the basketball motion recognition and basketball motion assessment tasks. The experimental results show that the proposed networks achieve competitive recognition results on the public dataset, and achieve high recognition and assessment accuracy on the constructed basketball skeleton dataset.

The following part of this paper is organized as follows: the related works are reviewed in Section 2; the basketball skeleton dataset is constructed in Section 3; the architecture of the proposed skeleton motion recognition network and skeleton motion assessment network are proposed in Section 4; the experiments are provided in Section 5; Section 6 is the conclusion.

2. RELATED WORKS AND ANALYSIS

As an important research content in the field of computer vision, human motion recognition has broad application prospects in intelligent monitoring, human-computer interaction, virtual reality and video analysis. At present, the research on human motion recognition can be roughly divided into video based human motion recognition (Qian et al. 2021) and skeleton based human motion recognition (Zhang et al. 2017).

- **Video based human motion recognition methods:** Include traditional methods and deep learning methods. With the application of deep learning in the field of computer vision and the availability of large-scale video data sets, the deep learning based research algorithms have achieved far better results than traditional methods in the field of motion recognition, and become the mainstream research direction in this field. Simonyan et al. (2014) proposed a dual flow network composed of spatial flow network and temporal flow network on the basis of 2D CNN, in which spatial flow network is used to model appearance features and temporal flow network is used to model motion features. Tran et al. (Tran et al. 2015) proposed C3D method, which used three-dimensional convolution to model spatial-temporal signals, and obtained a more compact feature representation than 2D CNN. Tran et al. (Tran et al. 2017) extended the C3D architecture to the deep residual network, and proposed the Res3D network. Carreira et al. (Carreira et al. 2017) used 3D convolution and pooling operations to expand the inception network, proposed the inflated 3D ConvNet (I3D), which adopted greater spatial-temporal resolution for the input of I3D network, and proposed a new method to initialize 3D CNN. However, video based methods require high cost computing resources to process pixel information in RGB images and optical streams. In the face of complex background conditions, the amount of computation is larger and the robustness is poor. With the rapid development of depth imaging technology, the Kinect depth sensor is used to obtain 3D human skeleton data, and the 3D space trajectory changes of main joint points are used to describe motions. In such manner, the calculation efficiency is high, and the background noise is robust. Therefore, skeleton based human motion recognition has attracted more attention from researchers.
- **Skeleton based human motion recognition/ Kinect body sensor based motion recognition:** The early skeleton motion recognition models are usually developed based on manually designed features, for example, the time series feature vector (Wang et al. 2012) can be formed according to the joint coordinate point position in the skeleton data sequence, or consider the relative positions between joint points, and form feature vectors (Wang et al. 2013) according to the posture changes, offsets and other information of different areas of the body. It is also possible to take into account the static characteristics of human posture and the dynamic shift characteristics of time series as a whole to form feature vectors (Yang et al. 2012; Vemulapalli et al. 2014; Fernando et al. 2015). Based on these features, traditional machine learning algorithms are used for classification and recognition. The disadvantage of these methods is that the ability of feature expression is limited. When the amount of data increases and the types of actions increase, the recognition accuracy is not high. With the rapid development of deep learning, the deep learning based skeleton motion recognition methods have gradually become mainstream, and can be divided into two categories: recurrent neural network (RNN) based methods, convolutional neural network (CNN) or graph convolutional network (GCN) based methods.

RNN is suitable for modeling data with strong temporal correlation, such as, in order to build a deeper network, Li et al. (Li et al. 2018) proposed an RNN model based on independent neurons, which improved the accuracy of skeleton motion recognition. In order to better express the context dependence of time series data and solve the gradient dispersion problem of traditional RNN, the long short term memory (LSTM) structure has been more widely used. For instance, the spatial temporary

LSTM (ST-LSTM) model (Liu 2016) and global context aware attention LSTM (GCA-LSTM) model (Liu 2018) proposed by Liu et al. However, RNN based methods are still difficult to deal with the simultaneous changes of skeleton data in spatial and temporal dimensions.

In the CNN based methods, the skeleton data sequence is converted into a set of semantic images by preset rules, and then processed by CNN model. For example, the three-dimensional coordinates of joint points can be directly mapped to the three channel pixel values of the image (Li et al. 2017), or selecting several reference joint points, calculate the relative position change from other joint points to the reference point, and consider the change of view angle to generate a series of continuous images (Ke et al. 2017; Ke et al. 2018). Compared with RNN, the CNN based methods have a certain improvement in recognition performance, but they still cannot fully use the natural internal information of skeleton data, and the inherent correlation between human joints cannot be well modeled and expressed. The good performance of graph convolution neural network in many fields provides a new idea for skeleton motion recognition. Since the skeleton is naturally constructed as a graph in non-Euclidean geometric space, Yan et al. (Yan et al. 2018) constructed a spatial-temporal skeleton map, and proposed spatial composite GCN (ST-GCN) model, which embedded the spatial-temporal relationship between human joints into the adjacency matrix of the skeleton map. In order to further explore the natural connection between joints and the dependence of joints in different degrees during the movement, Li et al. (Li et al. 2019) proposed actional-structural GCN (ASGCN) model based on dynamic connection and static structural connection. Zhang et al. (Zhang et al. 2020) used semantic information such as the type of joint (hand joint, shoulder joint, etc.) to guide the establishment of a more effective GCN model to fully express global and local features. However, the model that only depends on the preset topology of the graph is not flexible enough to deal with the diverse motion data. Shi et al. (Shi et al. 2019) proposed an adaptive graph convolution model to learn the individual data driving matrix in an end-to-end manner, which is a supplement to the graph topology adjacency matrix, and increases the data adaptability of GCN model. Chen et al. (2021) designed a double headed GCN model to effectively combine coarse and fine grain characteristics. Ma Li et al. (Ma et al. 2022) proposed a region association adaptive GCN model to capture the internal dependencies between non-physical joints. In recent years, the application of GCN methods has significantly improved the recognition rate of skeleton motion, and has become the mainstream method. However, most GCN methods are not deep enough to analyze the dynamic changes of skeleton motion data series in time domain, especially for the relationship between motion units with non-adjacent time steps, which is ignored due to information dilution.

The Transformer proposed by Vaswani et al. (Vaswani et al. 2017) has become the leading model in the NLP field, because it has achieved excellent results in processing very long sequences and parallelizing sentences, which LSTM and RNN do not have. Recently, some researchers have tried to introduce Transformer into the field of computer vision using only multiple head self-attention layers (Dosovitskiy et al. 2021), and have produced the most advanced results on image classification benchmarks (such as ImageNet). There are also methods (Ramachandran et al. 2019), to apply the Transformer model at the image pixel level, but this will cause a large computing cost, and the image must be down sampled or local attention must be used instead of global attention. At present, researchers have introduced Transformer networks into skeleton motion recognition tasks to model spatial/temporal long-distance dependence.

3. THE COLLECTION OF BASKETBALL SKELETON DATASET

At present, there have exist many large human skeleton datasets, such as NTU-RGB+D datasets, which are collected by Microsoft Kinect v2 sensor and contain 60 daily behaviors. However, these datasets are not collected specifically for basketball motion recognition, and it is difficult to directly use them in basketball course assessment. Besides, some existing basketball motion recognition datasets are not publicly available, which leads to our inability to directly use these data. In addition, many existing data sets contain not only the behavior of a single person, but also the interaction between people and objects. The dataset we want to collect only includes the behavior of a single person.

Based on the above requirements, this section will build the basketball skeleton dataset. The collected dataset consists of two parts: basketball motion recognition part and basketball skill assessment part. Specifically, referring to the construction process of NTU-RGB+D dataset, Kinect v2 sensor is also used to collect our skeleton dataset.

For the basketball motion recognition (Basketball-Motion) part, the basketball motion sequence is collected according to six motions: standing dribble, walking dribble, running dribble, jump shot, free throw, and layup. Finally, 300 video sequences are collected in the basketball motion recognition part after processing, and each type of motion contains 50 sequences, each sequence contains dozens to hundreds of 3D skeleton images, and the lens is guaranteed to be unobstructed, and the background is as free of interference as possible. In addition, in order to be used for technical assessment tasks, besides to the above 300 standard motion sequences, 120 non-standard motion video sequences (20 non-standard sequences for each type of motion) will also be collected.

The specific data collection process is as follows: After the Kinect v2 sensor is arranged, 50 senior students majoring in physical education are first selected to collect 6 types of standard motions at a fixed distance. A total of 300 standard motion skeleton video sequences are then obtained. After that, 20 amateurs imitated the above 6 types of motions respectively, and 120 non-standard motion video sequences were obtained, which were also converted into skeleton sequence. Figure 1 shows several skeleton images in a video sequence.

In order to mark the assessment task, 10 professionals were recruited to score each video sequence collected using the 5-point system. Specifically, the standard action video sequence is scored as 5 points, while other non-standard video sequences are scored respectively, with the highest score of 3 points and the lowest score of 1 point. The scoring of some samples is shown in Table 1.

As can be seen from Table 1, since most of the collected data are from beginners, the overall score is between 1.5 and 2.4.

4. THE SKELETON MOTION RECOGNITION NETWORK AND SKELETON MOTION ASSESSMENT NETWORK

4.1 ViT Based Skeleton Motion Recognition Network

In order to model the long-distance dependency between skeleton sequences, this part uses ViT as our backbone network. Let $X = \{x_1, x_2, \cdots, x_n\}$ be a skeleton sequence of n frames, where x_i denotes the i^{th} skeleton image, which has been reshaped to a vector. One naive approach to handle

Figure 1.
Some sample examples of the collected basketball skeleton dataset

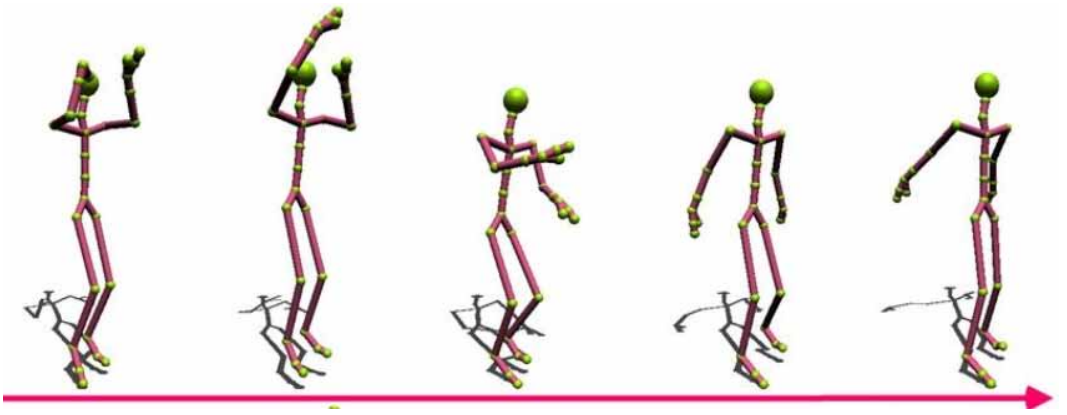


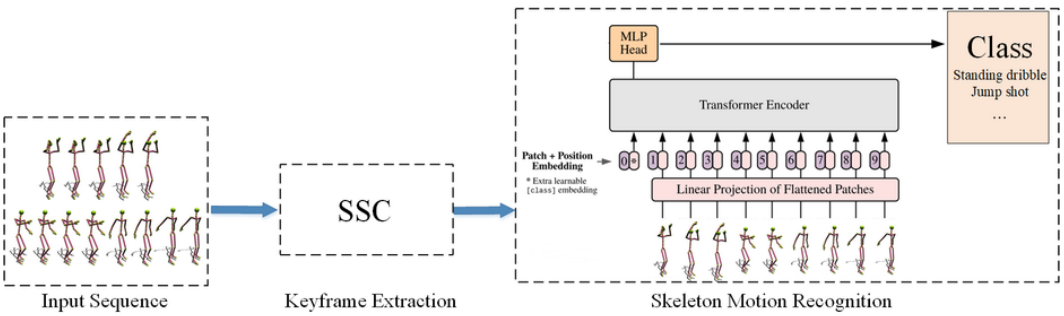
Table 1.
Data of participants' aesthetics assessment on basketball skill

ID	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	Average
1	3	3	3	1	2	1	3	2	3	3	2.4
2	2	3	2	3	3	2	2	1	1	1	2.0
3	2	3	1	3	2	3	3	2	2	3	2.4
4	3	2	2	2	3	2	2	2	2	2	2.2
5	2	2	1	1	1	1	1	2	2	2	1.5

the motion recognition task is to feed these sequences into ViT directly. However, there are usually dozens or even hundreds of frames in our sequence. If each frame skeleton is fed into the network as a patch in the ViT, the computing cost will increase dramatically. Therefore, in order to solve this problem, the keyframe extraction technology is considered. That is, before feeding sequences into ViT, we will first cluster sequence frames by using the sparse subspace clustering (SSC) (Elhamifar et al. 2009) algorithm, and take the cluster centers as the keyframes of each sequence, so the sequence with the original size of n frames become to $Y = \{y_1, y_2, \cdots, y_k\}$, where k denotes the number of keyframes. Thus, the flow chart of our proposed ViT based skeleton motion recognition network is shown in Figure 2. Specifically, our motion recognition network consists of two modules: keyframe extraction module, and basketball motion recognition module. The keyframe extraction module uses SSC to extraction keyframes from one sequence, and the basketball motion recognition module is a variant of ViT, in other words, we use the extracted keyframe as the patch.

- **Keyframe extraction module:** In this module, the SSC is used to extract keyframes. Suppose that $X = [x_1, x_2, \cdots, x_n] \in \mathbb{R}^{p \times n}$, where p denotes the feature dimension, n denotes the total number of skeleton frames. Since all frames in the sequence describe the same object or person, it can be considered that these samples are located in the same linear subspace, or they can be used as the basis vector for generating the subspace $span\{X\}$. In other words, any sample in X can be represented as a linear combination of all other frames. Thus,the self-representation can be formulated as $X = XC + E$, where $C = [c_1, c_2, \cdots, c_n] \in \mathbb{R}^{n \times n}$ denotes the self-representation coefficient matrix, and E is the representation error matrix (which usually can be ignored). In addition, a constraint term $diag(C) = 0$ is usually used to avoid trivial solution. Thus, the objective function of SSC can be formulated as:

Figure 2.
Flowchart of the proposed ViT based skeleton motion recognition network



$$\begin{aligned} &\min \left\| C \right\|_1 \\ &s.t.\ X = XC,\ \ diag(C) = 0 \end{aligned} \tag{1}$$

The l_1 norm is used to ensure that we use as few samples as possible to reconstruct/linearly represent the target frame. The optimization problem can be solved by many methods, such as the alternating direction method of multipliers (ADMM). After C is computed, we can construct an affinity matrix W :

$$W = \frac{1}{2}\left(\mid C \mid + \mid C \mid^T \right) \tag{2}$$

It will be used as the input of spectral clustering algorithm to get the final clustering results, and then we can get the k keyframes:

$$Y = \{y_1, y_2, \cdots, y_k\}$$

- **Basketball motion recognition module:** In this module, the k keyframes are used as the inputs of the Transformer. However, since Transformer needs a constant latent vector size D through all of layers, so a trainable linear projection is first used to project the p -dimension features to D dimensions:

$$z_0 = [x_{class}; y_1 E; y_2 E; \cdots; y_k E] + E_{pos} \tag{3}$$

where $E \in \mathbb{R}^{p \times D}$ is the projection matrix, $E_{pos} \in \mathbb{R}^{(k+1) \times D}$ denotes the position embedding, $x_{class} \in \mathbb{R}^D$ denotes the class token, which will be used for classification after it pass the whole network. z_0 denotes the output of the 0-layer, in other words, it denotes the input of the 1-layer. Since ViT consists of multiple multiheaded self-attention (MSA) blocks and before each block the layer-norm (LN) is applied, so we have the following outputs of all blocks:

$$z_l^{'} = MSA(LN(z_{l-1})) + z_{l-1},\ \ l = 1, \cdots, L \tag{4}$$

$$z_l = MLP(LN(z_l^{'})) + z_l^{'},\ \ l = 1, \cdots, L \tag{5}$$

$$y = LN(z_L^0) \tag{6}$$

where the subscript l denotes the l^{th} block. And y is the feature representation of the whole skeleton sequence, which will be used for classification (i.e., used as the input of the classification layer). MLP denotes multi-layer perceptron, which is essentially a full connection layer.

4.2 ViT Based Skeleton Motion Assessment Network

The flow chart of our proposed ViT based skeleton motion assessment network is shown in Figure 3. As shown in Figure 3, the assessment network consists of three parts, the keyframe extraction module, basketball motion recognition module, and the basketball motion assessment module, respectively. The first two modules are identical to the recognition network. In other words, we use the same network to learn the feature representation y of the entire skeleton sequence and obtain the category of each video sequence. After obtaining the motion class of the test sequence, you can bring it to the basketball motion assessment module. Since there are 6 motion categories, we need to train

6 branches (or sub-networks) in this module. In addition, because the number of training samples in each category is very limited, for each category, a 3-layer fully connected neural network is used to build the assessment classifiers. And the number of output neurons of each branch is 5, which respectively represents the probability that the input sequence belongs to scoring 1,2,3,4,5.

Finally, in order to better train our network, the EMD-based loss is used as our loss function (which is introduced in NIMA (Talebi et al. 2018):

$$EMD(p,\hat{p}) = \left(\frac{1}{N} \sum_{k=1}^{N} \left\| \sum_{i=1}^{k} p_{s_i} - \sum_{i=1}^{k} \hat{p}_{s_i} \right\|^r \right)^{1/r} \tag{7}$$

where p and $\hat{p}$ are the ground truth and estimated probability respectively. And here s_i denotes the i^{th} score bucket, $N = 5$ represents the total number of score buckets, r is a hyper-parameter.

5. EXPERIMENTAL RESULTS AND ANALYSIS

In order to verify the effectiveness of the proposed networks, we will perform experiments on the basketball motion recognition task and the basketball motion assessment task respectively.

For the basketball motion recognition tasks, the experiments will be performed on the public NTU-RGB+D dataset (Shahroudy et al. 2016), and the Basketball Skeleton dataset collected in this paper.

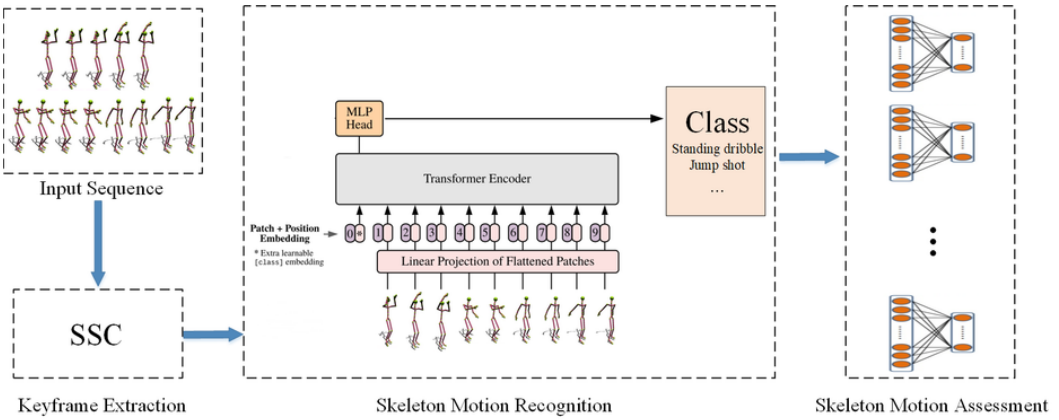
For the assessment task, we only perform experiments on the collected Basketball Skeleton dataset.

Our methods are implemented using Pytorch. The baseline weights are initialized by training on ImageNet, and the last fully-connected layer is randomly initialized. The weight and bias momentums are set to 0.9, and a dropout rate of 0.5 is applied on the last layer of the baseline network. The learning rate of the baseline network is set as 1e-6.

5.1 Performance on Basketball Motion Recognition Task

NTU-RGB+D dataset is one of the largest and most widely used indoor capture motion recognition datasets. It includes 56880 RGB videos, depth sequences, skeleton data and infrared frames sampled from RGB+D video samples. And the skeleton information is composed of 3D coordinates of 25 body joints, representing 60 different motion categories. The NTU-RGB+D dataset follows two

Figure 3.
Flowchart of the proposed ViT based skeleton motion assessment network



different evaluation benchmarks: CrossSub (X-Sub) and Cross View (X-View). In this paper, we only use the CrossSub (X-Sub) benchmark for testing, that is, the data set is divided into training set (40,320 videos) and verification set (16,560 videos), and the actors in the two subsets are different.

For the collected Basketball Skeleton dataset, the recognition part is used, that is, 80% of the 300 standard motions, i.e., 240 sequences, are used as the training set, and the remaining 60 sequences are used as the verification set. And for motion recognition tasks, the classification accuracy is used as a performance criterion.

Since the number of keyframes k is an important parameter, we will first verify the effect of parameter k. Here k is taken as 6, 9, 12, 24, 50. And the experimental results are shown in Figure 4.

It can be seen from this figure that when the number of keyframes vary, the recognition rates will also vary, this means that the parameter k has a certain impact on the recognition performance. More specifically, when k is very small (for example $k = 5$), the recognition rates on two datasets are about 84% and 92%. With the number of keyframe increase, the recognition performance also increases. However, we also notice that when $k > 24$, the recognition rate will decrease. In addition, the larger the k, the longer the calculation time. Hence, we think $k = 24$ is an appropriate choice for our recognition tasks.

Next, we set $k = 24$, and perform comparison experiments on NTU-RGB+D and Basketball Skeleton datasets. And the comparison results are shown in Table 2. Here, the comparison methods include C3D, I3D, Deep LSTM, and Co-ConvT (Shi et al. 2022). As shown in Table 2, we can observe that in all cases, our proposed recognition network achieves the best recognition rates, which demonstrate that our proposed method can effectively deal with the motion recognition task, especially the basketball motion recognition task. Specifically, the proposed recognition network achieves the best accuracy 90.3% on the public NTU-RGB+D dataset, which is about 2.2% higher than the state-of-the-art method. Besides, on our Basketball Skeleton dataset, we also find that our proposed network performs much better than the comparison methods, which means that our proposed method can effectively deal with the basketball motion recognition task.

Figure 4.
Average classification results of our proposed recognition network on NTU-RGB+D and Basketball Skeleton datasets with different *k* (%)

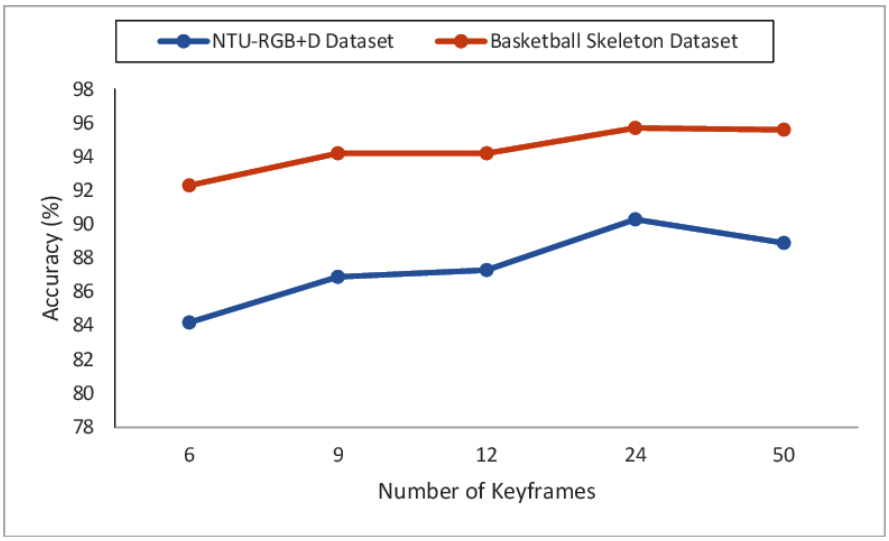


Table 2.
Average classification results of different methods on NTU-RGB+D and Basketball Skeleton datasets (%)

Dataset	NTU-RGB+D Dataset	Basketball Skeleton Dataset						
Methods		Motion 1	Motion 2	Motion 3	Motion 4	Motion 5	Motion 6	Average
C3D	85.3	92.3	94.3	93	90.9	90.3	95.7	92.8
I3D	86.2	93	94.6	93	91.2	90	96.0	93.0
Deep LSTM	86.0	94.2	95.3	92.8	91.7	90.9	98	93.8
Co-ConvT	88.1	95	96	94.8	92	92.8	96	94.4
Our	90.3	96	95.7	97.5	94	93.3	97.5	95.7

5.2 Performance on Basketball Motion Assessment Task

For the basketball motion assessment task, all 300+120 sequences were used. Specifically, 40 standard motions, 16 non-standard motions per class are used for training set, while the remaining 10+4 sequences per class are selected as testing set.

For the assessment task, the linear correlation coefficient (LCC), Spearman's rank correlation coefficient (SRCC) and EMD value are used as our performance criterion.

Similar to the above, we also first verify the effect of parameter k, and the experimental results are shown in Figure 5.

As shown in Figure 5, our method is not very sensitive to parameter k, and in most cases we find that our method has strong correlation with the ground truth. In addition, we also find that when $k = 24$, the proposed network performs the best, hence, in what follows, it is acceptable to set $k = 24$.

Next, we will compare the comparison experiments with some state-of-the-art assessment methods, such as NIMA, A-Lamp CNN (Ma et al. 2017). The comparison results on our collected dataset are shown in Table 3.

Figure 5.
Average performance of the proposed network on Basketball Skeleton dataset with different *k*

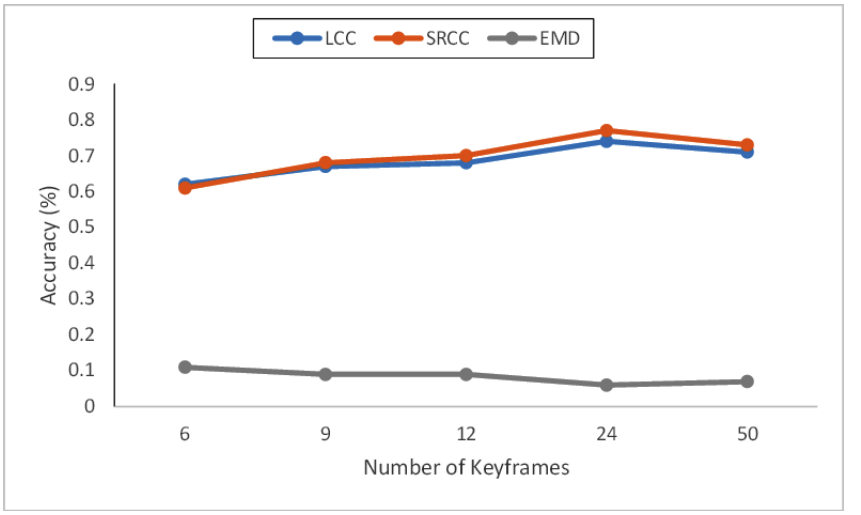


Table 3.
Performance of the proposed method in predicting the score of basketball skill compared to the state-of-the-art

Methods	LCC	SRCC	EMD
A-Lamp CNN	0.60	0.59	0.09
NIMA(Inception-v2)	0.67	0.63	0.07
Our	0.74	0.77	0.06

From Table 3, we find again that our proposed method performs better than the comparison method (NIMA and A-Lamp CNN) in all attributes, which can demonstrate the effectiveness of our ViT based skeleton motion assessment network. The results also show that the proposed method can be effectively used in the quantitative assessment of basketball courses.

6. CONCLUSION

In this paper, aiming at the problem of Kinect based basketball courses quantitative assessment, we first collect a new basketball skeleton dataset that consists of two parts: basketball motion recognition part and basketball skill assessment part. Then, the collected skeleton data is viewed as video sequence, and the new ViT based skeleton motion recognition network and ViT based skeleton motion assessment network are developed to perform motion recognition task and motion assessment task, respectively. In addition, in order to decrease the computing time, the SSC algorithm is used to extract keyframes. The experimental results on public dataset and our collected dataset show the superiority of our algorithms to some state-of-the-art methods. Our future research includes continuing to expand the basketball skill assessment dataset and exploring more intelligent basketball skill assessment methods.

REFERENCES

Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action recognition? A new model and the kinetics dataset. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724-4733. doi:10.1109/CVPR.2017.502

Chen, T., Zhou, D., & Wang, J. (2021). Learning multigranular spatio-temporal graph network for skeleton-based action recognition. *Proceedings of ACM International Conference on Multimedia*, 4334-4342.

Dosovitskiy, A., Beyer, L., & Kolesnikov, A. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *The 9th International Conference on Learning Representations*.

Elhamifar, E., & Vidal, R. (2009). Sparse subspace clustering. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2790–2797.

Fernando, B., Gavves, E., & Oramas, J. (2015). Modeling video evolution for action recognition. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2015*, 5378–5387.

Gao, X., Niu, S., Wei, D., Liu, X., Wang, T., Zhu, F., & Sun, Q. (2022). *Joint metric learning-based class-specific representation for image set classification. IEEE Transactions on Neural Networks and Learning Systems*.

Ke, Q., & Bennamoun, M. (2017). A new representation of skeleton sequences for 3D action recognition. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2017*, 3288–3297.

Ke, Q., & Bennamoun, M. (2018). Learning clip representations for skeleton-based 3D action recognition. *IEEE Transactions on Image Processing*, *27*(6), 2842–2855.

Li, B., Dai, Y., & Cheng, X. (2017). Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. *Proceedings of IEEE International Conference on Multimedia and Expo Workshops*, 601-604.

Li, M., Chen, S., & Chen, X. (2019). Actional-structural graph convolutional networks for skeleton-based action recognition. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019*, 3595–3603.

Li, S., Li, W., & Cook, C. (2018). Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2018*, 5457–5466.

Liu, J., Shahroudy, A., & Xu, D. (2016). Spatio-temporal LSTM with trust gates for 3D human action recognition. *European Conference on Computer Vision*, 816-833.

Liu, J., Wang, G., & Duan, L. (2018). Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Transactions on Image Processing*, *27*(4), 1586–1599.

Ma, L., Zheng, S., & Niu, B. (2022). Action recognition method on regional association adaptive graph convolution. *Journal of Frontiers of Computer Science and Technology*, *16*(4), 898–908.

Ma, S., Liu, J., & Chen, C. (2017). A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qian, H., Yi, J., & Fu, Y. (2021). Review of human action recognition based on deep learning. *Journal of Frontiers of Computer Science and Technology*, *15*(3), 438–455.

Ramachandran, P., Parmar, N., & Vaswani, A. (2019). Stand-alone self-attention in vision models. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 7.

Shahroudy, A., & Liu, J., & Ng, T. T. (2016). NTU RGB+D: A large scale dataset for 3D human activity analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 1010–1019.

Shi, L., Zhang, Y., & Cheng, J. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019*, 12026–12035.

Shi, Y., & Zhu, M. (2022). Collaborative Convolutional Transformer Network for Skeleton-based Action Recognition. *Dianzi Yu Xinxi Xuebao*, 1–9.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Proceedings of the Advances in Neural Information Processing Systems*, 568-576.

Talebi, H., & Milanfar, P. (2018). NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, *27*(8), 3998–4011.

Tran, D., Bourdev, L., & Fergus, R. (2015). Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the 15th IEEE International Conference on Computer Vision*, 4489-4497. doi:10.1109/ICCV.2015.510

Tran, D., Ray, J., & Shou, Z. (2017). *ConvNet architecture search for spatiotemporal feature learning*. https://arxiv.org/abs/1708.05038

Vaswani, A., Shazeer, N., & Parmar, N. (2017). Attention is all you need. *The 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Vemulapalli, R., Arrate, F., & Chellappa, R. (2014). Human action recognition by representing 3D skeletons as points in a Lie group. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2014*, 588–595. doi:10.1109/CVPR.2014.82

Wang, C., Wang, Y., & Yuille, A. (2013). An approach to posebased action recognition. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 915-922.

Wang, J., Liu, Z., & Wu, Y. (2012). Mining actionlet ensemble for action recognition with depth cameras. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1290–1297. doi:10.1109/CVPR.2012.6247813

Wu, J., Ma, M., Jiang, M., & Luo, K. (2018). Sitting posture intention judgmental based on Kinect and its application. *Computer Applications and Software*, *35*(10), 194–199.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *2018*, 7444–7452.

Yang, X., & Tian, Y. (2012). EigenJoints-based action recognition using naïve-Bayes-nearest-neighbor. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 14-19.

Zhang, P., Lan, C., & Zeng, W. (2020). Semantics-guided neural networks for efficient skeleton-based human action recognition. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2020*, 1112–1121.

Zhang, Y., Chang, F., & Liu, H. (2017). Action recognition based on 3D skeleton. *Tien Tzu Hsueh Pao*, *45*(4), 906–911.

Zheng, W., Chen, S., Fu, Z., Zhu, F., Yan, H., & Yang, J. (2022). Feature selection boosted by unselected features. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(9), 4562–4574. doi:10.1109/TNNLS.2021.3058172 PMID:33646957

Zheng, W., Xu, C., Yang, J., Gao, J., & Zhu, F. (2018). Low-rank structure preserving for unsupervised feature selection. *Neurocomputing*, *314*, 360–370. doi:10.1016/j.neucom.2018.06.010

Zhu, F., Gao, J., Yang, J., & Ye, N. (2022). Neighborhood linear discriminant analysis. *Pattern Recognition*, *123*, 108422. doi:10.1016/j.patcog.2021.108422

Zhu, F., Ning, Y., Chen, X., Zhao, Y., & Gang, Y. (2021, April 1). On removing potential redundant constraints for SVOR learning. *Applied Soft Computing*, *102*, 106941.